RESEARCH PROPOSAL

"A STATISTICAL ANALYSIS OF E-HEALTHCARE DATA USING MACHIN LEARNING TECHNIQUES"

A synopsis submitted to

SWAMI RAMANAND TEERTH MARATHWADA UNIVERSITY, NANDED, MAHARASHTRA

For the degree of

DOCTOR OF PHILOSOPHY

IN

STATISTICS

UNDER THE FACULTY OF SCIENCE AND TECHNOLOGY

By

Miss. Ashwini Digambar Wakale

UNDER THE GUIDANCE OF

Dr. Vikas Shivaji Jadhav

Assistant Professor,

Department of Statistics

Sanjeevanee Mahavidyalaya, Chapoli Tal. Chakur Dist. Latur

JANUARY – 2024

"A STATISTICAL ANALYSIS OF E-HEALTHCARE DATA USING MACHINE LEARNING TECHNIQUES"

1) INTRODUCTION :

Rising healthcare costs have been a major issue for developed nations [1].This cost has been a major concern in the India, and therefore early detection is important. In light of the rapid advancement of biotechnology, and an era of big data generated for healthcare by mainly EHR (Electronic Health Records) in various structures, it is increasingly more important to intelligently use this information to make sense of hidden patterns, detect abnormalities, and predict healthcare diseases.

Artificial intelligence has certainly made computers smarter. Machine learning which is a subset of artificial intelligence plays an important role in mining large datasets and extracting valuable knowledge from them [3,6]. Training a machine appropriately with proper train data set, the machine's algorithm can learn patterns and therefore detect any abnormalities in the initial stage of a disease which can help patients save overall cost and time. This research study will examine the opportunities of machine learning and data mining in the healthcare industry especially in various diseases, how early diagnosis can minimize healthcare costs; and how data generated by EHR can provide insights for medical professionals in terms of detecting abnormalities for potential chronic diseases. We begin by providing a brief research background, followed by the problem statement, research questions, objectives, and the organization of this culminating experience research.

Nowadays, there is huge amount of data being collected and stored in databases everywhere across the world. Today, the healthcare industry which is one of the largest industries throughout the world includes medical industries having the large amounts of health-related and medical-related data. It also includes thousands of hospitals, clinics and other types of facilities that provide primary, secondary and tertiary levels of care. The delivery of healthcare service is the most visible part of any healthcare system, both to users and the general public. Accurate, early and error-free diagnosis and treatment given to patients has been a major issue highlighted in medical service. Quality service in health care field implies diagnosing patients correctly and administering treatments that are effective. To achieve a correct and cost-effective treatment, a system can be developed to fulfill the task.

Today, healthcare systems use EHR system to manage and store their healthcare or patient data [8]. EHR (Electronic Health Record) is digital repository of Patients' Case

Papers, medical history, treatments, diagnosis, medications, allergies, immunization dates, etc. These systems generate huge amounts of medical data which may contain their computer-readable entries like Magnetic Resonance Imaging (MRIs), signals like Electrocardiography (ECG), clinical information like blood sugar, blood pressure, cholesterol levels, etc. as well as the physician's interpretation. These data may comprise a lot of invaluable information and knowledge hidden in such databases, which can be used to support effective clinical decision-making. But the discovery of hidden patterns and knowledge from such databases often goes unexploited.

Machine learning and Data mining can improve the management level of hospital information, saving time and cost. The discovered knowledge can then be applied to the medical data to increase the working efficiency and help medical practitioners in their decision-making. This provides near-endless opportunities for symptom trend detection, earlier detection of illness, and DNA trend analysis, etc. thus increasing efficiency in healthcare (E. N. Ekwonwune et al. 2022). Also, we can compare efficacy, usefulness, consistency, cost effectiveness, etc. for the traditional treatment methods to treat any particular disease like, Ayurveda, Unani, Siddha, Allopath, Homeopathy, etc. Data mining EHRs for comparative studies between above mentioned different medicinal streams can contribute significantly to the broader understanding of Healthcare Practices. It can help bridge the gap between traditional and modern medicine, facilitate evidence based decision making, and potentially lead to more comprehensive and personalized healthcare solutions.

2) REVIEW OF LITERATURE AND DEVELOPMENT IN THE SUBJECT:

In recent years, data mining and machine learning techniques has emerged as a powerful tool in healthcare industry, offering various advancements and developments. Bardhwaj et al., (2017), Shailaja et al., (2018), Sun et al., (2019), and Lee & Yoon, (2017) studied a broad overview of machine learning techniques used in healthcare for various diseases. They provided insights into the potential value of medical big data that can be used for clinical decision support, diagnostics, treatment decisions, fraud detection, and prevention. They briefly summarized the nine-step data mining process along with focusing on why efficient decision support was required by the healthcare system. The results from their experiment showed that machine learning models can be used for the early diagnosis of diseases.

A review by Rishi Dubey et al., (2015) studied different machine learning algorithms for the prediction of heart disease. Their study concluded that Neural Network is an efficient

technique for heart disease prediction. Ashish Chhabbi et al., (2016) used a dataset collected from UCI repository to perform different data mining techniques to predict heart disease. They applied K-means algorithm and Naïve Bayes and their results revealed that tuning the number of clusters of the k-means algorithm gave better results than the default K-means.

Boshra Baharami et al., (2015) evaluated various classification methods such as, Decision Tree, K-Nearest Neighbors (k-NN), and SMO (used to train Support Vector Machines). On their dataset, they used feature selection techniques to only select the important attributes and achieved the highest accuracy of 83.732% with Decision Trees. Mustafa et al., (2018) proposed an ensemble approach for better prediction by combining five classifiers. Their work included SVM, ANN, Naïve Bayes, Regression analysis and Random Forest their goal was to predict and diagnose cardiovascular disease.

The literature review reveals emerging and advanced machine learning and data mining algorithms involved in predicting heart diseases. It is evident from the above literature review that data mining algorithms have effectively predicted health diseases. The trustworthiness of the model for predicting heart diseases with different risk factors is a high concern, however, SVM, Naïve Bayes, Decision Trees, Bagging and Boosting, and Random Forest have achieved reliable results in the diagnosis of heart disease (Jan et al., 2018).

In the above literature review, many different data mining prediction models have been introduced such as SVM, Naïve Bayes, Decision Trees, Bagging and Boosting, and Random Forest for heart disease. The models using these algorithms to predict heart disease produced very high accuracy. Therefore, we move forward to review the literature that aligns with our study objective which is how machine learning algorithms can be used in the healthcare data prediction.

Clinical Decision Support Systems (CDSS): Mohammed Badawy et al. (2023) surveyed on Healthcare predictive analytics using machine learning and deep learning techniques integration of data mining into CDSS provides healthcare providers with real-time insights and recommendations based on historical data. This assists in clinical decision-making, suggesting optimal treatment paths and reducing medical errors.

Healthcare Fraud Detection: Matthew Herland et al. (2018) worked on Big Data fraud detection using multiple medicare data sourses. Data mining techniques are employed to identify anomalies or irregularities in billing, insurance claims, and healthcare records. This aids in detecting fraudulent activities, reducing financial losses, and ensuring compliance with regulations.

Overall, machine learning advancements in healthcare continue to revolutionize the industry by enabling more informed decision-making, improving patient care, facilitating targeted interventions, and driving innovations in medical research and healthcare delivery.

3) PROBLEM STATEMENT:

Modern information technology tools and techniques such as AI, Machine Learning and Data Mining could help support healthcare system, data-driven approaches to enhance patient care, optimize resource allocation, and improve healthcare delivery systems, we aim to contextualize the significance of this study within the broader landscape of healthcare analytics, emphasizing the need for innovative methodologies to extract actionable insights from complex healthcare datasets [6].

4) OBJECTIVES OF THE RESEARCH:

The objective of this research study is to explore how machine learning algorithms can be use in the E-healthcare data prediction.

1. To identify patterns and trends in healthcare data that can be used to make predictions.
2. To identify potential drug interactions and adverse reactions.
3. To improve the accuracy and efficacy of diagnosis and treatment.
4. To reduce healthcare costs by predicting and preventing costly medical events.
5. To identify and target high-risk populations for preventive care.
6. To improve the accuracy of medical decision-making.

5) RESEARCH METHODOLOGY:

Machine learning plays very important role in e-healthcare data analysis and deriving knowledge from such EHR databases. Statistical analysis and machine learning techniques help healthcare professionals in the diagnosis of disease.

There are several different methodologies to approach this research study like Classification, Regression Trees, Decision Trees, Support Vector Machines, and Association Rules etc. Some common methods used in machine learning or data mining include: [3].

1. Classification: Classification involves categorizing data into predefined classes or categories based on input features. Algorithms like decision trees, support vector machines (SVM), k-nearest neighbors (KNN), and neural networks are commonly used for classification tasks.

2. Clustering: Clustering aims to group similar data points together based on their inherent characteristics or similarities. K-means clustering, hierarchical clustering and density-based clustering (DBSCAN) are popular methods for clustering analysis.

3. Regression Analysis: Regression analysis predicts numerical values or continuous outcomes based on input variables. Linear regression, multiple regression, logistic regression, and polynomial regression are used to model relationships between variables.

Some of the parametric and non-parametric trend models are also used to this study for checking the trend behavior of the data series. The models along with their equations are given below [14].

a) Linear Regression Model: A linear model is one which all the parameters appear linearly it is formulated as:

$$Y = a + bX + e$$

Where, Y= Dependent variable, X = Independent variable,
a = Intercept, b = Slope and
e = Random error.

b) Multiple Regression Model:

$$Yi = \beta 0 + \beta 1x1 + \beta 2\, x2 + \ldots\ldots + \beta nxn + \in$$

Where, for i = n Observations, Yi = Dependent variable,
X_i = Explanatory variable, $\beta 0$ = Intercept (constant term),
βn = Slop coefficient for each explanatory variable,
$\in$ = The model's error term (also known as the residuals)

c) Logistic Regression Model: Logistic regression is a parametric model that calculates the probability for an event to happen based on the data.

$$f\ (x) = \frac{1}{1 + e - x}$$

Where, e = base of natural logarithms.

4. Time Series Analysis: Time series analysis deals with sequential data points captured over time. Methods like ARIMA (Autoregressive Integrated Moving Average), exponential smoothing, and LSTM (Long Short-Term Memory) networks are used for forecasting and trend analysis [15].

a) ARIMA Model: An autoregressive (AR) integrated moving average – ARIMA model is generalization of a simple autoregressive moving average – ARIMA model. Both of these models are used to forecast or predict future points in the time-series data. ARIMA is a form of regression analysis that indicates the strength of dependent variable relative to other changing variables. It is an extrapolation method that requires only historical time series data on the variable under study.

5. Decision Trees:

Decision tree is similar to flow chart in which every non-leaf node denotes a test on a particular attribute and every branch represents an outcome of the test. Root node is the topmost node in the decision tree. For example, with the help of readmission tree, we can decide whether a patient needs to be readmitted or not. Using Decision Tree, a decision maker can choose the best alternative and traversal from root to leaf indicates unique class separation based on maximum information gain. Decision tree are self-explanatory and easy to follow. Set of rules can also be constructed with the help of decision tree. Decision Tree can be considered as nonparametric method because there is no need to make assumptions regarding distribution of space and structure of classifier [2, 6].

6. **Support Vector Machines:** A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes. Support Vector Machine algorithms have been used effectively to predict healthcare dataset.
7. **Association Rule Mining:** It is also referred to as market basket analysis, where data is analyzed to identify relationship between entities. It is used in various domains for discovering patterns and relationships within data. Algorithms like a priori and FP-growth are often used to generate association rules from datasets.

6) IMPORTANCE OF STUDY:

To handle machine learning techniques are becoming increasingly important in healthcare. These techniques can be used to analyze large amounts of data to identify patterns and trends that can be used to predict outcomes and inform decision-making. By leveraging machine learning techniques, healthcare providers can gain insights into patient health, identify potential risks, and develop more effective treatments [2].

Additionally, these techniques can be used to improve the accuracy of diagnosis and prognosis, as well as to identify potential areas of improvement in healthcare delivery. By utilizing machine learning and data mining techniques, healthcare providers can gain a better understanding of their patients and the healthcare system as a whole, leading to improved patient outcomes and cost savings.

EHRs contain vast amounts of patient-related information. In this study, machine learning and data mining tools efficiently handle and process these large datasets, enabling the extraction of valuable insights. This will give algorithms to detect hidden patterns, trends, and relationships within EHR data. This capability assists in uncovering crucial information relevant to patient care, disease trends, or treatment effectiveness. These tools facilitate

predictive modelling. They help in forecasting patient outcomes, disease risks, or treatment responses based on patterns identified in EHRs, enabling proactive and preventive measures.

It will help healthcare organizations optimize resource allocation, reduce costs, and enhance operational efficiency, improving overall healthcare delivery [6]. This study will contribute to advancements in medical science and foster innovative healthcare solutions. Also data mining tools play indispensable role in EHR data analytics as they efficiently handle large datasets, unveil patterns, support predictive analytics, enable personalized medicine, enhance decision-making, drive efficiency, foster innovation, and ensure ethical data utilization in healthcare research and practice.

7) PROPOSED WORK PLAN:

- ✓ **First Year:**
  1. Assortment of topic: Study and background of research topic.
  2. To identifying review of literature through various national and international journals.
  3. Data collection and pre-processing of data.
  4. Detailed study and try to publish a review paper concerned topic.
- ✓ **Second Year:**
  1. Study of different software packages and to develop statistical model for healthcare data using machine learning techniques.
  2. Selection of one of the standard datasets for application healthcare data using machine learning algorithms.
  3. Prepare a comprehensive report detailing the analysis, key findings, limitations, and recommendations.
  4. Detailed study and statistical analysis of the results obtained (Model selection and development).
  5. The publications will be communicated based on the findings of the research.
- ✓ **Third Year:**
  1. Publication of papers in various national, international journals and conferences.
  2. Writing of final thesis and submission to the University.

8) REFRENCES :

1. Dadgostar, P; (2019), antimicrobial resistance: Implications and costs. Infection and Drug Resistance, Volume 12, 3903–3910. https://doi.org/10.2147/idr.s234610.
2. Bhardwaj, R., Nambiar, A. R., & Dutta, D. (2017), A study of machine learning in Healthcare. *IEEE 41st Annual Computer Software and Applications Conference* (COMPSAC). https://doi.org/10.1109/compsac.2017.164.
3. Shailaja, K., Seetharamulu, B., & Jabbar, M. A. (2018), Machine learning in healthcare: A Review; *Second International Conference on Electronics, Communication and Aerospace Technology* (ICECA). https://doi.org/10.1109/iceca.2018.8474918.
4. Lee, C. H., & Yoon, H.-J. (2017), Medical Big Data: Promise and challenges. Kidney Research and Clinical Practice, 36 (1), 3–11. https://doi.org/10.23876/j.krcp.2017.36.1.3
5. Rishi Dubey, and Santosh Chandrakar, Aug, (2015), "Review on Hybrid Data Mining Techniques for the Diagnosis of Heart Diseases in Medical Ground", Vol. 5, Issue 8, ISSN: 2249-555X, pp. 715-718.
6. Ashish Chhabbi,Lakhan Ahuja,Sahil Ahir, and Y. K. Sharma,19 March (2016),"Heart Disease Prediction Using Machine Learning Techniques", *International Journal of Research in Advent Technology*,E-ISSN:2321-9637,Special Issue National Conference "NCPC-2016", pp. 104-106.
7. Boshra Bahrami, and Mirsaeid Hosseini Shirvani, February (2015), "Prediction and Diagnosis of Heart Disease by Data Mining Techniques", *Journal of Multidisciplinary Engineering Science and Technology (JMEST),* ISSN: 3159- 0040, Vol. 2, Issue 2, pp. 164-168.
8. Zhu, Y., (2020), "Clustering Electronic Health Record Data for Identifying Patient Subpopulations: A Systematic Review." *Harvard Data Science Review*.
9. E. N. Ekwonwune1, C. I. Ubochi1, A. E. Duroha, (2022), "Data Mining as a Technique for Healthcare Approach" *Int. J. Communications, Network and System Sciences,* 15, 149-165 https://www.scirp.org/journal/ijcns ISSN Online: 1913-3723 ISSN Print: 1913-3715.
10. Soysal, E., et al., (2019), "Medication Side Effect Discovery Using Patient-Generated Health Data from Social Media: A Systematic Review." *Journal of the American Medical Informatics Association.*
11. Matthew Herland, Taghi M. Khoshgoftaar and Richard A. Bauder, (2018), "Big Data fraud detection using multiple medicare data sources", *Journal of Big Data,* https://doi.org/10.1186/s40537-18 38-3.

12. Matthew Herland, Taghi M Khoshgoftaar and Randall Wald, (2014), “A review of data mining using big data in health informatics”, *Journal of Big Data,* http://www.journalofbigdata.com/content/1/1/2.
13. Mohammed Badawy, Nagy Ramadan and Hesham Ahmed Hefny, (2023), “Healthcare predictive analytics using machine learning and deep learning techniques: a survey”, *Journal of Electrical Systems and Inf Technol*, https://doi.org/10.1186/s43067-023-00108-y.
14. D. C. Montgomery, E. A. Peck, G. G. Vining, (2012), "Introduction to Linear Regression Analysis", 5th edition, *Wiley Publication*.
15. D. C. Montgomery, C.L Gennings, M. Kulachi, (2010),"Introduction to Time Series Analysis and Forecasting", *A John Wiley & Sons, INC., Publication*.

Ms. Ashwini Digambar Wakale
(Research Student)

Dr. Vikas Shivaji Jadhav
(Research Guide)